Thésaurus distributionnels pour la recherche d'information et vice-versa

Vincent Claveau, Ewa Kijak

Thésaurus distributionnels pour la recherche d'information et vice-versa

Vincent Claveau* — **Ewa Kijak****

* *IRISA - CNRS* ** *IRISA - Univ. Rennes 1*
Campus de Beaulieu, 35042 Rennes
{vincent.claveau,ewa.kijak}@irisa.fr

RÉSUMÉ. *Les thésaurus distributionnels sont utiles à de nombreuses tâches du Traitement Automatique des Langues. Dans cet article, nous abordons les problèmes de leur construction et de leur évaluation sous l'angle de la recherche d'information. Deux contributions sont proposées. D'une part, en poursuite des travaux initiés par (Claveau* et al.*, 2014), nous montrons comment les techniques de RI peuvent être utilisées avec succès pour construire ces thésaurus. Au moyen d'une évaluation directe par comparaison avec des lexiques de référence et au travers de plusieurs expérimentations, nous montrons que les résultats obtenus par certains modèles de RI dépassent les performances des systèmes état-de-l'art. D'autre part, nous utilisons la RI comme cadre applicatif pour proposer une évaluation indirecte des thésaurus produits. Là encore, cette évaluation valide l'approche. Mais surtout, elle permet de mettre en regard les performances obtenues sur cette tâche avec celles des évaluations directes utilisées dans la littérature. Les différences constatées remettent en cause en partie ces pratiques d'évaluation.*

ABSTRACT. *Distributional thesauri are useful in many tasks of Natural Language Processing. In this paper, we address the problem of building and evaluating such thesauri with the help of Information Retrieval concepts. Two main contributions are proposed. First, in the continuation of the work of (Claveau* et al.*, 2014), we show how IR tools and concepts can be used with success to build thesaurus. Through several experiments and by evaluating directly the results with reference lexicons, we show that some IR models outperform state-of-the-art systems. Secondly, we use IR as an applicative framework to indirectly evaluate the generated thesaurus. Here again, this task-based evaluation validate the IR approach used to build the thesaurus. Moreover, it allows us to compare these results with those from the direct evaluation framework used in the literature. The observed differences question these evaluation habits.*

MOTS-CLÉS : *thésaurus distributionnels, sémantique distributionnelle, construction de lexique, modèles de RI, évaluation directe, évaluation par tâche, extension de requêtes.*

KEYWORDS: *distributional thesaurus, distributional semantics, lexicon generation, IR models, direct evaluation, task-based evaluation, query expansion.*

1. Introduction

La sémantique distributionnelle a pour objet de construire des thésaurus (ou lexiques) automatiquement à partir de corpus de textes. Pour une entrée donnée (ie. un mot donné), ces thésaurus recensent des mots sémantiquement proches en s'appuyant sur l'hypothèse qu'ils partagent une distribution similaire au mot d'entrée. En pratique, cette hypothèse distributionnelle est mise en œuvre simplement : deux mots seront considérés proches s'ils partagent des contextes similaires. Ces contextes sont typiquement les mots cooccurrants dans une fenêtre restreinte autour du mot examiné, ou les mots liés syntaxiquement à celui-ci.

L'évaluation de ces thésaurus reste un point crucial pour juger de la qualité des méthodes de construction employées. Une approche communément utilisée est de comparer le thésaurus produit à un ou plusieurs lexiques de référence. Cette évaluation, qualifiée d'intrinsèque, a pour avantage d'être directe et simple puisqu'elle permet d'estimer la qualité et la complétude du thésaurus produit. Cependant, elle repose sur des lexiques de référence dont la complétude, la qualité, ou tout simplement la disponibilité pour le domaine traité ne sont pas assurés.

Dans cet article, nous proposons d'examiner ces deux aspects – la construction et l'évaluation des thésaurus distributionnels – sous l'angle de la recherche d'information, utilisée à la fois comme technique et comme usage. Concernant la construction, des travaux récents (Claveau *et al.*, 2014) ont montré que des systèmes de RI pouvaient avantageusement être utilisés pour mettre en œuvre cette analyse distributionnelle.
Nous proposons dans cet article d'explorer cette approche RI de la construction des thésaurus en examinant l'intérêt de différents modèles classiques de RI en les comparant à l'état de l'art.
Concernant l'évaluation, nous proposons une évaluation extrinsèque des thésaurus produits dans une tâche de RI classique. Cela nous permet de mettre en regard ces résultats avec ceux obtenus par évaluation intrinsèque et donc de juger de la pertinence de ces scénarios d'évaluation.

Après un état-de-l'art (section suivante), l'article aborde ces deux contributions successivement : les aspects relatifs à la construction des thésaurus sont présentés en section 3, ceux portant sur l'évaluation par RI sont en section 4. Nous présentons enfin quelques conclusions et perspectives sur ce travail dans la dernière section.

2. État de l'art

2.1. *Construction de thésaurus distributionnels*

La construction de thésaurus distributionnels a fait l'objet de nombreuses études depuis les travaux pionniers de (Grefenstette, 1994) et (Lin, 1998). Toutes reposent sur l'hypothèse distributionnelle de (Firth, 1957) que l'on résume par sa formule célèbre : *"You should know a word by the company it keeps"*. On considère donc que chaque mot est caractérisé sémantiquement par l'ensemble des contextes dans lesquels

il apparaît. Pour un mot en entrée d'un thésaurus, des mots partageant des similarités de contextes sont proposés ; on les appelle voisins sémantiques par la suite. La nature du lien sémantique entre une entrée et ses voisins est variable ; ils peuvent être des synonymes de l'entrée, des hyperonymes, des hyponymes ou d'autres types de liens sémantiques (Budanitsky et Hirst, 2006 ; Adam *et al.*, 2013, pour une discussion)). Ces liens sémantiques, même s'ils sont très divers, sont néanmoins utiles pour de nombreuses applications liées au Traitement Automatique des Langues. Cela explique que ce champ de recherche soit encore très actif, avec des contributions portant sur différents aspects liés à la construction de ces thésaurus.

Tout d'abord, différentes pistes sur ce qui peut être considéré comme contexte distributionnel ont été explorées. On distingue ainsi les contextes graphiques des contextes syntaxiques. Les premiers sont simplement les mots apparaissant autour du mot étudié. Les seconds sont les mots recteurs ou dépendants syntaxiques du mot examiné. La seconde approche est souvent considérée comme plus précise, mais elle repose bien sûr sur une analyse syntaxique préalable qui n'est pas toujours disponible et peut même être source d'erreurs.

Les connexions entre sémantique distributionnelle et RI sont nombreuses. Plusieurs chercheurs ont par exemple utilisé des moteurs de recherche pour collecter des informations de co-occurrences ou des contextes sur le Web (Turney, 2001 ; Bollegala *et al.*, 2007 ; Sahami et Heilman, 2006 ; Ruiz-Casado *et al.*, 2005). Les représentations vectorielles des contextes sont également souvent utilisées de différentes manières (Turney et Pantel, 2010), mais sans lien avec les systèmes de pondérations et les fonctions de pertinence classiques de la RI (à l'exception de (Vechtomova et Robertson, 2012) dans un cadre un peu différent de similarité entre entités nommées). Plusieurs travaux se sont pourtant penchés sur la pondération des contextes pour obtenir de meilleurs voisins. Par exemple, (Broda *et al.*, 2009) a proposé de ne pas considérer directement les poids des contextes, mais les rangs pour s'affranchir de l'influence des fonctions de pondérations. D'autres ont proposé des méthodes d'amorçage (*bootstrap*) pour modifier les poids des contextes d'un mot en prenant déjà en compte ses voisins sémantiques (Zhitomirsky-Geffet et Dagan, 2009 ; Yamamoto et Asakura, 2010). Par ailleurs, beaucoup de travaux se sont basés sur le fait que la représentation "traditionnelle" des contextes distributionnels est très creuse et redondante, comme l'a illustré (Hagiwara *et al.*, 2006). Dans ce contexte, plusieurs méthodes de réduction de la dimension ont été testées : depuis l'analyse sémantique latente (Landauer et Dumais, 1997b ; Padó et Lapata, 2007 ; Van de Cruys *et al.*, 2011), jusqu'au *Random Indexing* (Sahlgren, 2001), en passant par la factorisation par matrices non négatives (Van de Cruys, 2010).

Récemment, (Claveau *et al.*, 2014) ont proposé d'identifier plus complètement le processus de recherche de voisins distributionnels comme un problème de recherche documentaire classique. L'ensemble des contextes d'un mot peut en effet être représenté comme un document ou une requête, ce qui permet de trouver facilement les mots proches, ou plus exactement les ensembles de contextes proches. Bien que partageant de nombreux points communs avec des travaux de l'état de l'art, cette façon

simple de poser le problème de la construction des thésaurus distributionnels offre des pistes intéressantes et un outillage facilement accessible. C'est cette approche que nous reprenons dans le cadre de cet article ; nous la décrivons plus en détail dans la section 3.1.

2.2. *Évaluation des thésaurus distributionnels*

Comme nous l'avons dit précédemment, l'évaluation des thésaurus produits se fait soit de manière intrinsèque, en les comparant à une ressource de référence, soit de manière extrinsèque, au travers de leur utilisation dans une tâche précise.

Dans le cas de l'évaluation intrinsèque, il faut disposer de lexiques de référence. Il est alors simple de calculer rappel, précision ou toute autre mesure de qualité du lexique produit. Cette approche a été utilisée pour évaluer de nombreux travaux. Parmi les lexiques régulièrement utilisés comme références, on peut citer WordSim 353 (Gabrilovich et Markovitch, 2007), ou ceux utilisés par (Ferret, 2013) qui exploitent des ressources plus larges, à savoir les synonymes de WordNet 3.0 (Miller, 1990) et le thésaurus Moby (Ward, 1996). Ce sont ces deux derniers lexiques que nous utilisons nous aussi pour l'évaluation intrinsèque ; voir ci-après pour une présentation. D'autres ressources ne sont pas directement des lexiques, mais des jeux de données permettant une évaluation directe, comme le jeu de synonymes du TOEFL (Landauer et Dumais, 1997a) ou l'ensemble de relations sémantique BLESS (Baroni et Lenci, 2011).

L'évaluation directe séduit par sa simplicité, mais pose la question de l'adéquation des lexiques utilisés comme références. Plusieurs recherches ont donc proposé des évaluations indirectes au travers d'une tâche. La plus connue est la tâche de substitution lexicale mise en œuvre à SemEval 2007 (McCarthy et Navigli, 2009). Étant donné un mot dans une phrase, le but est de remplacer ce mot par un de ses voisins supposés et de vérifier que cela n'altère pas le sens de la phrase. Les résultats obtenus sont ensuite comparés aux substitutions proposées par des humains. Cette tâche va donc privilégier les synonymes exacts au détriment des autres types de relations sémantiques. L'évaluation de thésaurus distributionnels par des tâches de RI n'a pas, à notre connaissance, été explorée. Bien sûr, l'utilisation d'informations que l'on peut qualifier de distributionnelles dans un cadre de RI a fait l'objet de plusieurs travaux (Besançon *et al.*, 1999 ; Billhardt *et al.*, 2002) qui se prolongent de nos jours par les travaux sur les représentations lexicales apprises par réseaux de neurones (Huang *et al.*, 2012 ; Mikolov *et al.*, 2013). Il s'agit dans tous les cas de tirer parti des similarités de contextes entre mots pour améliorer la représentation des documents et/ou la fonction de pertinence RSV. Cependant, ces travaux ne dissocient pas le processus de création du thésaurus distributionnel du processus de RI, ce qui rend l'évaluation de l'apport des informations distributionnelles seules impossible. Dans notre cas, l'évaluation extrinsèque par RI que nous proposons (cf. section 4) repose simplement sur l'utilisation des voisins sémantiques pour étendre des requêtes ; le reste du système de recherche d'information est standard. Cela doit nous permettre de juger au mieux de la qualité des thésaurus produits.

3. Modèles de RI pour l'analyse distributionnelle

3.1. *Principes et matériel*

Comme nous l'avons expliqué en introduction, le problème de la construction d'un lexique distributionnel peut être vu comme un problème de recherche de documents similaires et peut donc être mis-en-œuvre avec des techniques de RI. Dans ce cadre, pour un mot donné, ses contextes dans un corpus sont collectés et rassemblés. C'est cet ensemble de contextes qui forme un document. Construire une entrée du lexique, c'est-à-dire trouver les mots proches au sens distributionnel d'un mot w_i, revient alors à trouver les documents (contextes) proches du document représentant les contextes de w_i.

Le données que nous utilisons pour nos expériences de construction sont celles utilisées dans plusieurs travaux. Cela va nous permettre de comparer nos résultats à ceux publiés. Le corpus utilisé pour collecter les contextes est le corpus AQUAINT-2 ; il est composé de d'articles de presse en anglais et compte 380 millions de mots. Les mots que nous considérons pour entrées de notre lexique sont les noms communs apparaissant au moins 10 fois dans le corpus, soit 25 000 noms différents. Les contextes de toutes occurrences de ces mots sont donc collectés ; dans les expériences rapportées ci-dessous, on considère les deux mots à droite et deux mots à gauche du nom visé, en gardant leur position. Par exemple, dans l'extrait : "... all forms of restriction on freedom of expression, threats ...", les mots restriction-2, on-1, of+1, expression+2 sont ajoutés à l'ensemble des contextes freedom.

Comme nous l'avons évoqué précédemment, nous utilisons conjointement WordNet (WN) et Moby pour l'évaluation intrinsèque des thésaurus produits. Ces deux ressources offrent des caractéristiques complémentaires : WN recense des liens sémantiques forts (synonymes ou quasi-synonymes) alors que Moby recense une plus grande variété de liens (hyperonymes, méronymes, co-hyponymie...). Une description détaillée des liens considérés par ces ressources est donnée dans (Ferret, 2013 ; Claveau *et al.*, 2014). Ainsi, WN propose en moyenne 3 voisins pour 10 473 des noms du corpus AQUAINT-2 et Moby 50 voisins en moyenne pour 9 216 noms. Combinées, ces deux ressources couvrent 12 243 noms du corpus avec 38 voisins en moyenne. Le nombre de noms dans les listes de référence et la variété des relations sémantiques considérées font de ces données un jeu d'évaluation très complet par rapport à d'autres *benchmarks* parfois utilisés tels que WordSim 353 (Gabrilovich et Markovitch, 2007).

3.2. *Test des modèles de RI*

Le tableau 1 présente les résultats obtenus par différents systèmes de construction de thésaurus, appliqués au corpus Aquaint. Les mesures de performances utilisées pour comparer les thésaurus produits à la référence WordNet+Moby sont classiquement la précision à différents seuils, la MAP et la R-précision, exprimés en pour-

centage, moyennés sur les 12 243 noms de la référence WN+Moby et exprimés en pourcentage.

Méthode	MAP	R-Prec	P@1	P@5	P@10	P@50	P@100
Ferret 2013 *base*	5.6	7.7	22.5	14.1	10.8	5.3	3.8
Ferret 2013 *best rerank*	6.1	8.4	24.8	15.4	11.7	5.7	3.8
Ferret 2014 *synt*	7.9	10.7	29.4	18.9	14.6	7.3	5.2
Hellinger	2.45	2.89	9.73	6.28	5.31	4.12	3.30
TF-IDF	5.40	7.28	21.73	13.74	9.59	5.17	3.49
TF-IDF ajusté	7.09	9.02	24.68	15.13	11.55	5.96	4.31
Okapi-BM25	6.72	8.41	24.82	14.65	10.85	5.16	3.66
Okapi-BM25 ajusté	8.97	10.94	31.05	18.44	13.76	6.46	4.54
LSI dim=50	1.62	2.86	5.00	4.12	3.76	2.78	2.35
LSI dim=500	4.37	6.27	16.00	10.76	8.78	4.61	3.45
LSI dim=1000	5.06	6.87	21.09	13.20	9.96	5.39	4.02
LSI dim=2000	5.11	6.86	23.11	14.34	10.78	5.12	3.72
LDA dim=500	0.60	1.25	2.17	2.21	1.90	1.29	1.13
RP dim=500	5.66	6.48	27.3	12.85	8.67	3.04	1.86
RP dim=2000	5.90	7.04	27.13	13.71	8.94	3.21	1.96
LM Dirichlet $\mu = 25$	6.52	7.56	23.46	11.88	8.16	2.99	1.89
LM Dirichlet $\mu = 250$	6.56	7.43	23.08	12.31	8.17	2.77	1.73
LM Dirichlet $\mu = 2500$	5.83	6.77	23.28	12.06	8.00	2.98	1.81
LM Hiemstra $\lambda = 0.45$	5.41	6.79	25.09	12.07	8.17	3.05	1.90
LM Hiemstra $\lambda = 0.65$	8.10	8.98	27.06	13.35	9.25	3.41	2.13
LM Hiemstra $\lambda = 0.85$	7.06	7.88	25.28	12.44	8.41	3.04	1.89
LM Hiemstra $\lambda = 0.95$	6.49	7.64	27.21	13.62	9.17	3.28	2.06

Tableau 1 – Performances des modèles de RI pour la construction des thésaurus distributionnels sur la référence WN+Moby

À des fins de comparaison, nous rapportons les résultats obtenus dans les mêmes conditions expérimentales avec une approche état-de-l'art notée *base* exploitant une similarité cosinus avec une pondération par information mutuelle (Ferret, 2013), une version avec apprentissage (*rerank*) pour réordonnancer les voisins (Ferret, 2013), et une version (*synt*) reposant non plus sur des contextes graphiques, mais syntaxiques (Ferret, 2014). Nous rapportons également les résultats des systèmes déjà testés par (Claveau *et al.*, 2014), qui reposent sur la similarité d'Hellinger (Escoffier, 1978 ; Domengès et Volle, 1979), un TF-IDF/cosinus, et Okapi-BM-25 (Robertson *et al.*, 1998). Ces mêmes auteurs proposent une version dite ajustée de la similarité Okapi-BM25, dans laquelle l'influence de la taille du document est renforcée, en prenant $b = 1$, et en mettant l'IDF au carré pour donner plus d'importance aux mots de contexte plus discriminants. Nous appliquons également cette stratégie pour obtenir une version ajustée du TF-IDF/cosinus en prenant l'IDF au carré.

En plus de ces modèles, nous testons des systèmes probabilistes par modèles de langues (notés LM), avec lissage de Dirichlet (avec différentes valeurs du paramètre μ), et lissage à la Hiemstra (lissage avec les probabilités d'apparition des mots dans toute la collection ; avec différentes valeur de λ). Nous testons également des tech-

niques de réductions de dimensions (LSI, LDA, Random Projections (RP)), avec différents nombres de dimensions.

Ces modèles de RI, très classiques, ne sont pas détaillés plus avant ici ; le lecteur intéressé trouvera les notions et détails utiles dans les références citées ou des ouvrages généralistes (Manning *et al.*, 2008 ; Boughanem et Savoy, 2008, par exemple).

On observe tout d'abord la difficulté de la tâche puisque dans tous les cas, les précisions relevées sont très faibles selon cette évaluation intrinsèque. La comparaison avec les lexiques de référence conduit donc à une conclusion très sévère quant à la qualité supposée des thésaurus produits. On note tout de même que certains modèles de RI fonctionnent particulièrement bien par rapport à l'état de l'art, comme les modèles basés sur Okapi, ou les modèles de langues.

Les techniques de réduction de dimension donnent des résultats d'autant plus limités que le nombre de dimensions considérées est petit. Ce résultat faible est en ligne avec certaines conclusions de travaux précédents (Van de Cruys, 2010). Le fait d'agréger en une seule dimension des mots différents est donc préjudiciable pour bien distinguer les voisins sémantiques, autrement dit, l'apparition de certains mots de contexte bien précis est un indicateur fort pour juger de la proximité sémantique des mots. Cela est d'ailleurs confirmé par le fait qu'au sein d'une même famille de modèle, les paramètrages menant aux meilleurs résultats sont ceux qui donnent plus de poids aux mots discriminants : IDF au carré pour Okapi, faible lissage pour les modèles de langue (μ et λ relativement petits).

3.3. *Analyse par fréquence*

Certains auteurs ont remarqué que la fréquence des mots dont on essaie de trouver les voisins a une grande influence sur la qualité finale (Ferret, 2013). Plus ils sont fréquents, plus on a de contextes pour les décrire et meilleurs sont les résultats avec les méthodes état-de-l'art. On se propose donc de vérifier si l'emploi de méthodes issues de la RI amène la même observation. Pour cela, on reprend le cadre expérimental précédent et le modèle Okapi ajusté, mais on distingue les résultats selon la fréquence des mots-entrées : les mots ayant les plus hautes fréquences (>1000), ceux avec les fréquence les plus basses (<100) et le tiers restant avec des fréquences moyennes. Ces résultats sont présentés dans le tableau 2. Là encore nous indiquons les résultats état-de-l'art de (Ferret, 2013) pour comparaison.

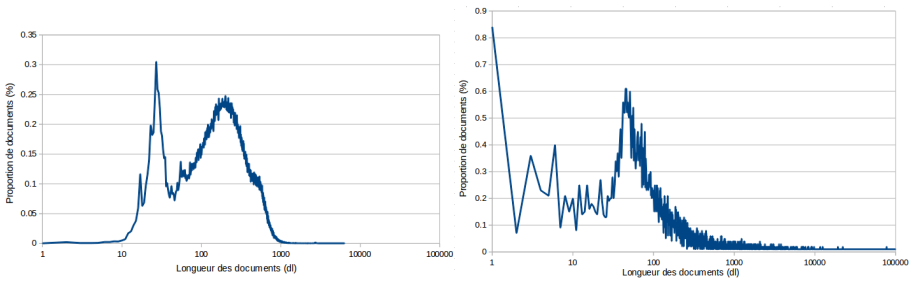
Il apparaît que l'approche par RI a un comportement bien plus stable selon les fréquences que le système état-de-l'art de (Ferret, 2013). En particulier, l'approche RI assure des résultats de bonne qualité pour les mots faiblement fréquents. La fréquence des mots étant directement liée à la taille des ensembles de contextes, cela indique l'importance de normalisation en fonction de la taille des documents dans l'approche RI.

Fréq.	Méthode	MAP	R-Prec	P@1	P@5	P@10	P@50	P@100
élevée	Ferret 2013 *base*	6.5	11.0	41.3	26.8	20.8	-	7.3
	Okapi ajusté	7.21	10.73	39.78	24.8	19.31	9.16	5.99
moyenne	Ferret 2013 *base*	7.4	9.3	20.9	12.3	9.3	-	3.2
	Okapi ajusté	9.85	11.32	30.58	16.19	11.85	5.19	3.55
basse	Ferret 2013 *base*	2.4	2.1	3.3	1.7	1.5	-	0.7
	Okapi ajusté	6.93	6.79	9.88	4.83	3.84	1.97	1.49

Tableau 2 – Performances pour la construction des thésaurus distributionnels sur la référence WN+Moby selon les fréquences des mots visés

3.4. *Limites de l'analogie avec la RI*

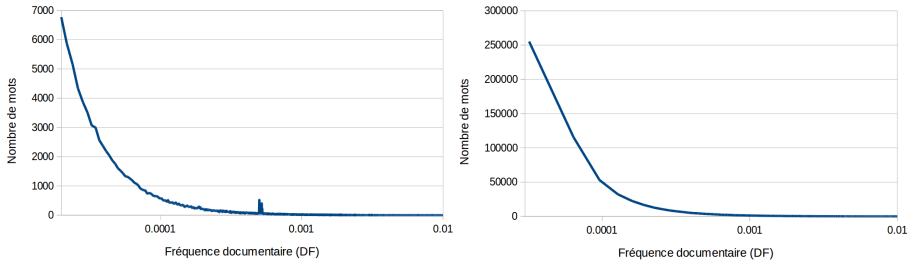
L'analogie entre recherche de document similaire et recherche de voisins distributionnels apporte de très bons résultats, mais il convient cependant de pointer certaines limites de cette analogie. En effet, les ensembles de contextes, qui sont considérés comme des documents, ont des propriétés sensiblement différentes des documents réels. Pour illustrer cela, nous produisons en figure 1 la distribution des tailles de documents standard (ce sont ceux du corpus Aquaint, c'est-à-dire des articles de journaux) et de celles des ensembles de contextes. On y observe un éventail de taille beaucoup plus important dans le cas des ensembles de contextes. Il semble donc important dans les fonctions de similarités utilisées de prendre en compte ce besoin de normalisation accrue selon la longueur des documents.



(a) (b)

Figure 1 – Distributions des tailles des documents dans le cadre standard (a) et des ensembles de contextes (b) ; échelle log.

La distribution des mots est également assez différente de ce que l'on trouve dans une vraie collection de documents. Cela est illustré en figure 2 dans laquelle on donne la distribution des fréquences documentaires (DF), en se comparant là encore avec le corpus Aquaint original. Les mots apparaissent en général dans beaucoup plus de contextes que c'est le cas pour de vrais document. Par exemple, le nombre de mots



(a) (b)

Figure 2 – Distributions des fréquences documentaires (DF) dans le cadre standard (a) et des ensembles de contextes (b) ; échelle log.

apparaissant dans 1 document sur 10 000 (DF=0.0001) est près de 100 fois plus élevé que pour de vrais documents. Comme nous l'avions vérifié expérimentalement, ce phénomène milite pour une prise en compte spécifique de cette distribution dans les modèles (à travers les lissages dans les modèles de langue ou l'IDF dans les modèles vectoriels par exemple, ou de nouveaux schémas de pondérations).

4. Évaluation dans un cadre de RI

Pour évaluer l'apport des thésaurus distributionnels dans une tâche classique de RI, nous nous plaçons dans un cadre d'extension de requêtes. Pour chaque nom de la requête, les mots associés dans le thésaurus distributionnel sont ajoutés à celle-ci. Nous décrivons ci-dessous notre contexte expérimental, puis les résultats obtenus. Nous proposons ensuite de mettre en regard les résultats obtenus par cette évaluation indirecte avec les résultats de l'évaluation intrinsèque que nous avons utilisée précédemment.

4.1. *Contexte expérimental*

La collection de RI que nous utilisons est celle développée pour le projet Tipster et utilisée dans le cadre de TREC. Elle contient plus de 170 000 documents et cinquante requêtes. Ces requêtes sont composées de plusieurs champs (la requête à proprement parler, un champ narratif détaillant les critères de pertinence) ; dans les expériences rapportées ci-dessous, nous n'utilisons que le champ requête. Cette collection est particulièrement adaptée puisqu'elle est composée de documents en anglais de même nature que le corpus Aquaint (articles du *Wall Street Journal*) à partir duquel le thésaurus distributionnel a été construit.

Le système de recherche d'information que nous utilisons est Indri (Metzler et Croft, 2004 ; Strohman *et al.*, 2005), connu pour offrir des performances état-de-l'art.

Ce système probabiliste implémente une combinaison de modèle de langue (Ponte et Croft, 1998) et de réseaux d'inférence (Turtle et Croft, 1991). Dans les expériences rapportées ci-dessous, nous l'utilisons avec des réglages standard, à savoir un lissage de Dirichlet ($\mu = 2500$). Dans notre cas, ce système de RI offre l'avantage supplémentaire de disposer d'un langage de requête complexe qui nous permet d'inclure les mots du thésaurus distributionnel en exploitant au mieux le modèle par réseau d'inférence à l'aide de l'opérateur dédié '`#syn`' qui permet d'agréger les comptes des mots considérés comme synonymes (voir la documentation d'Indri pour plus de détails). Pour supprimer les effets de flexions (pluriel) sur les résultats, les formes pluriel et singulier des noms de la requêtes sont ajoutées, que ce soit dans la requête non étendue avec les synonymes ou celles étendues par les voisins sémantiques.

Les performances pour cette tâche de RI sont également classiquement mesurées en précision à différents seuils (P@x), R-prec, MAP. L'évaluation du lexique consiste donc en la comparaison des résultats obtenus avec ou sans extension, que nous mesurons en gain relatif de précision, de MAP... Nous indiquons également la moyenne des gains d'AP par requête, notée AvgGainAP (à ne pas confondre avec le gain de MAP, qui est le gain calculé sur les moyennes des AP par requête). Les résultats non statistiquement significatifs (Wilcoxon et t-test avec $p < 0.05$) sont en italiques.

4.2. *Résultats d'extension*

Le tableau 3 présente les gains de performance obtenus en étendant les requêtes avec les mots collectés dans les thésaurus. Nous choisissons le lexique ayant obtenu les meilleurs résultats : celui construit avec la méthode Okapi ajustée. Puisque ce lexique ordonne les voisins par proximité avec le mot-entrée, on teste différents scénarios : pour chaque mot de la requête, on ne garde que ses 5, 10 ou 50 plus proches voisins. À des fins de comparaison, on indique aussi les résultats obtenus en étendant avec les lexiques de référence WN seul et WN+Moby. Voici un exemple de requête, avec sa forme non-étendue et sa forme étendue (Okapi ajusté top 5) utilisant les opérateurs de réseau d'inférence d'Indri :

– requête : `coping with overcrowded prisons`

– forme normale : `#combine( coping with overcrowded #syn( prisons prison ) )`

– forme étendue : `#combine( coping with overcrowded #syn( prisons prison inmate inmates jail jails detention detentions prisoner prisoners detainee detainees ) )`

On note tout d'abord que quelque soit le lexique utilisé, l'extension de requête apporte un gain significatif de performance. Comme beaucoup de travaux depuis, cela contredit au passage les conclusions de (Voorhees, 1994) sur l'absence d'intérêt à utiliser WN pour étendre des requêtes. Le fait le plus notable est cependant les excellentes performances du lexique construit automatiquement, qui dépassent même celles des lexiques de référence. Alors que sa précision sur les 10 premiers voisins a

Extension	MAP	AvgGainAP	R-Prec	P@5	P@10	P@50	P@100
Sans	21.78	-	30.93	92.80	89.40	79.60	70.48
avec WN	+12.44	+36.3	+7.01	+4.31	+7.16	+7.60	+10.87
avec WN+M	+11.00	+28.33	+7.78	*+3.02*	+5.37	+6.53	+9.17
avec Okapi-BM25 ajusté top 5	+13.14	+29.99	+11.17	+3.45	+5.15	+9.40	+12.43
avec Okapi-BM25 ajusté top 10	+13.80	+24.36	+9.58	*+2.16*	+4.03	+5.58	+8.26
avec Okapi-BM25 ajusté top 50	+10.02	+17.99	+8.82	+3.45	+3.36	+3.72	+5.36

Tableau 3 – Gains relatifs de performance (%) par extension de requête selon le lexique utilisé

été évaluée à moins de 14 % en section 3, ce lexique produit des extensions obtenant le meilleur gain en MAP. La moyenne des gains d'AP (AvgGainAP) apporte également des informations intéressantes : celle-ci est maximale avec WN, qui offre donc une amélioration stable (concernant beaucoup de requêtes) grâce au fait qu'il ajoute à la requête principalement des voisins très proches sémantiquement (synonymes exacts). Cette stabilité diminue avec les autres lexiques, et est la plus basse avec les extensions par les 50 plus proches voisins du lexique généré par le modèle Okapi ajusté. Comme la MAP reste globalement bonne, cela indique que seules certaines requêtes bénéficient d'un gain absolu important.

4.3. *Évaluation intrinsèque vs. évaluation extrinsèque*

Les résultats de l'expérience précédente soulèvent des questions sur la cohérence entre les résultats de l'évaluation intrinsèque et ceux de l'évaluation extrinsèque. Le gain de précision entre deux méthodes de construction de thésaurus, même s'il est jugé statistiquement significatif, est-il sensible en RI ? Pour cela on complète les résultats précédents avec la figure 3 qui rapporte les résultats de différents modèles de RI à la tâche d'extension (avec les 10 premiers voisins) selon leur P@10 de l'évaluation directe. Il en ressort que la précision mesurée avec l'évaluation directe est liée aux gains mesurés dans la mesure où l'ordre est bien respecté : la meilleure P@10 à l'évaluation directe obtient le meilleur gain de MAP à la tâche de RI, etc. Mais la corrélation n'est pas linéaire comme on pourrait s'y attendre. D'autre part, des différences statistiquement significatives lors de l'évaluation directe (comme entre TFIDF ajusté et Oakpi ajusté) ne se traduisent pas forcément par des différences statistiquement significatives à la tâche d'extension. Parmi les faux positifs de l'évaluation directe (mots détectés comme proches mais absents dans les lexiques de référence), certains semblent plus ou moins néfastes pour étendre les requêtes.

Il est alors intéressant d'examiner plus précisément l'effet de ces faux positifs. On examine de nouveau l'évolution des performances sur la tâche de RI en fonction de la qualité des listes de voisins utilisées pour étendre les requêtes, mais cette fois-ci, des listes de voisins plus ou moins bruitées sont générées à partir des thésaurus de référence en remplaçant des voisins par des mots choisis aléatoirement dans le vo-

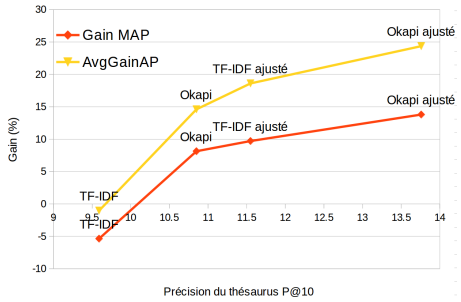


Figure 3 – Gain en MAP et AvgGainAP de différents modèles selon leur précision @10 lors de l'évaluation intrinsèque

cabulaire. On peut ainsi produire des listes de voisins avec une précision variable et contrôlée, dont on évalue les performances pour étendre les requêtes comme précédemment. La figure 4 montre l'évolution de la MAP et de l'AvgGainAP en faisant varier ainsi la précision des listes de référence WN seul et WN+Moby. On indique pour comparaison les scores obtenus avec les top 5, 10 et 50 du lexique Okapi ajusté.

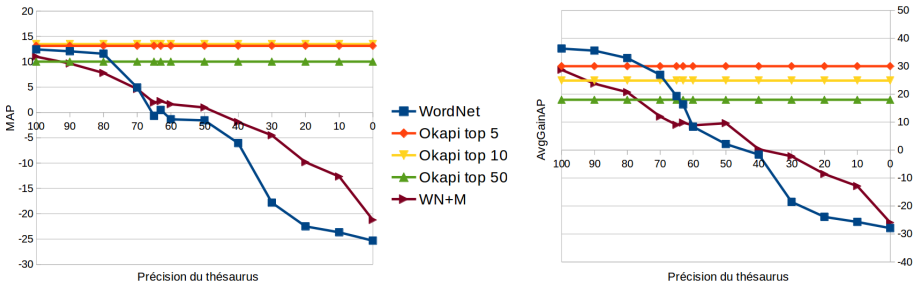


Figure 4 – Gain en MAP (gauche) et AvgGainAP (droite) selon la précision contrôlée artificiellement des thésaurus utilisés pour étendre les requêtes

Comme attendu, les deux mesures de performance chutent lorsque la précision des listes diminue. Il faut une précision des listes inférieure à 50 % pour rendre les gains de performance nuls, et en deçà, les extensions dégradent les résultats. Il y a donc bien une corrélation entre la précision des listes mesurée par évaluation directe et les performances pour l'extension de requête, du moins lorsque que les faux positifs sont pris au hasard. Mais dans le cas du lexique que nous avons généré, les performances obtenues sont comparables à des listes de précision entre 70 et 100 % selon les cas, alors que la précision mesurée par évaluation intrinsèque variait entre 10 et 20 %. Plus que la sévérité de l'évaluation intrinsèque, cela souligne la faiblesse de la démarche

Extension avec Okapi-BM25 ajusté	MAP	AvgGainAP	R-Prec	P@5	P@10	P@50	P@100
top 10 sauf WN	+11.80	+21.60	+8.37	*+2.16*	+3.58	+5.08	+6.87
top 10 sauf WN+M	+9.36	+19.22	+6.41	*+3.02*	+3.36	*+3.17*	+5.73

Tableau 4 – Gains relatifs de performance (%) par extension de requête avec les voisins jugés faux positifs

qui repose sur des références incomplètes : certains voisins, jugés comme faux positifs car non listés par les références sont en réalité de bons candidats.

Pour illustrer ce dernier point, nous rapportons dans le tableau 4 les performances obtenues par le lexique Okapi ajusté en étendant de nouveau les requêtes avec les 10 premiers voisins de chaque nom, mais en excluant ceux qui sont listés comme voisins dans WN ou WN+Moby. Autrement dit, on ne garde que les voisins jugés comme faux positifs par l'évaluation intrinsèque. Il apparaît clairement que ces faux positifs sont bien liés sémantiquement à l'entrée. Pour le mot `prison` de la requête précédente, parmi les 10 premiers voisins, ceux absents de WN+Moby sont : `sentence, abuse, detainee, guard, custody, defendant, inmate, prisoner`. Ils semblent effectivement bien liés sémantiquement à *prison*.

5. Conclusion

Dans cet article, nous avons exploré l'utilisation de la recherche d'information à la fois pour construire et pour évaluer des thésaurus distributionnels. Nous avons d'une part utilisé les modèles de similarités développés en RI sur les contextes des mots, ce qui nous permet, pour un mot donné, de trouver ceux partageant une similarité contextuelle, et donc sémantique. D'autre part, la recherche d'information, à travers la tâche classique de recherche de documents par requête, nous offre un cadre applicatif permettant une évaluation indirecte des thésaurus.

De ces travaux, deux conclusions majeures se dégagent. En étendant les propositions de (Claveau *et al.*, 2014), nous avons confirmé le bien-fondé de l'approche RI pour la construction des thésaurus sémantiques. Nous avons en particulier montré l'importance de la prise en compte des mots discriminants dans différents modèles (au travers de pondérations spécifiques pour l'IDF ou par le lissage). Nous avons également souligné l'avantage des modèles RI par rapport aux méthodes classiques en particulier sur les mots avec peu d'occurrences. Mais nous avons également souligné les limites de l'analogie en RI et sémantique distributionnelle : les ensembles de contextes ont des propriétés statistiques (taille, fréquence d'apparition des mots...) très différentes de 'vrais' documents. Cela milite pour l'établissement de fonctions de pondération et de pertinence adaptées à cette réalité et ouvre donc des voies d'amélioration possibles. D'autres perspectives sur ce point concernent l'utilisation de techniques

récentes de RI pour la construction des thésaurus (*learning to rank*, représentations continues...).

L'autre conclusion majeure de cet article porte sur la fiabilité de l'évaluation intrinsèque. En montrant que les thésaurus obtenus offrent des résultats au moins aussi bons que les listes de référence servant à l'évaluation intrinsèque, nous remettons en perspective beaucoup de conclusions de travaux précédents. Les faibles résultats obtenus aux évaluations intrinsèques ne se traduisent pas dans ce cadre applicatif d'extension de requête. Il convient bien sûr de nuancer cette conclusion, qui ne porte que sur le cadre applicatif testé : la tâche et la mise en œuvre que nous utilisons (avec les opérateurs de croyances d'Indri) permet d'avoir des liens sémantiques relativement distants dans les listes de voisins servant d'extensions sans que cela ne dégrade trop les résultats. D'autres tâches, comme la substitution lexicale, plus centrée sur la synonymie exacte, pourraient donner d'autres résultats. Une perspective intéressante serait ainsi de mesurer la corrélation entre les scores d'évaluation intrinsèque et extrinsèque dans différentes tâches pour mieux aider à choisir les méthodes de construction les plus adaptées selon la tâche finale visée.

6. Bibliographie

Adam C., Fabre C., Muller P., « Évaluer et améliorer une ressource distributionnelle : protocole d'annotation de liens sémantiques en contexte », *TAL*, vol. 54, n° 1, p. 71-97, 2013.

Baroni M., Lenci A., « How we BLESSed distributional semantic evaluation. », *Workshop on GEometrical Models of Natural Language Semantics*, p. 1-10, 2011.

Besançon R., Rajman M., Chappelier J.-C., « Textual Similarities based on a Distributional Approach », *in Proceedings of the Tenth International Workshop on Database and Expert Systems Applications (DEXA'99)*, p. 180-184, 1999.

Billhardt H., Borrajo D., Maojo V., « A Context Vector Model for Information Retrieval », *J. Am. Soc. Inf. Sci. Technol.*, vol. 53, n° 3, p. 236-249, February, 2002.

Bollegala D., Matsuo Y., Ishizuka M., « Measuring semantic similarity between words using web search engines », *Proceedings of WWW'2007*, 2007.

Boughanem M., Savoy J. (eds), *Recherche d'information : états des lieux et perspectives*, Hermès Science, avril, 2008.

Broda B., Piasecki M., Szpakowicz S., « Rank-Based Transformation in Measuring Semantic Relatedness », *22^{nd} Canadian Conference on Artificial Intelligence*, p. 187-190, 2009.

Budanitsky A., Hirst G., « Evaluating WordNet-based Measures of Lexical Semantic Relatedness », *Computational Linguistics*, vol. 32, n° 1, p. 13-47, 2006.

Claveau V., Kijak E., Ferret O., « Improving distributional thesauri by exploring the graph of neighbors », *International Conference on Computational Linguistics, COLING 2014*, Dublin, Irlande, August, 2014.

Domengès D., Volle M., « Analyse factorielle sphérique : une exploration », *Annales de l'INSEE*, vol. 35, p. 3-83, 1979.

Escoffier B., « Analyse factorielle et distances répondant au principe d'équivalence distributionnelle », *Revue de statistique appliquée*, vol. 26, n° 4, p. 29-37, 1978.

Ferret O., « Identifying Bad Semantic Neighbors for Improving Distributional Thesauri », *51^{st} Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria, p. 561-571, 2013.

Ferret O., « Typing relations in distributional thesauri », *in* N. Gala, R. Rapp, G. Bel (eds), *Advances in Language Production, Cognition and the Lexicon*, Springer, 2014.

Firth J. R., *Studies in Linguistic Analysis*, Blackwell, Oxford, chapter A synopsis of linguistic theory 1930-1955, p. 1-32, 1957.

Gabrilovich E., Markovitch S., « Computing semantic relatedness using wikipedia-based explicit semantic analysis », *20^{th} International Joint Conference on Artificial Intelligence (IJCAI 2007)*, p. 6-12, 2007.

Grefenstette G., *Explorations in automatic thesaurus discovery*, Kluwer Academic Publishers, 1994.

Hagiwara M., Ogawa Y., Toyama K., « Selection of Effective Contextual Information for Automatic Synonym Acquisition », *21^{st} International Conference on Computational Linguistics and 44^{th} Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, Sydney, Australia, p. 353-360, 2006.

Huang E. H., Socher R., Manning C. D., Ng A. Y., « Improving word representations via global context and multiple word prototypes », *50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*, p. 873-882, 2012.

Landauer T., Dumais S., « A solution to Plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge », *Psychological Review*, vol. 104, n° 2, p. 211-240, 1997a.

Landauer T. K., Dumais S. T., « A solution to Plato's problem : the latent semantic analysis theory of acquisition, induction, and representation of knowledge », *Psychological review*, vol. 104, n° 2, p. 211-240, 1997b.

Lin D., « Automatic retrieval and clustering of similar words », *17^{th} International Conference on Computational Linguistics and 36^{th} Annual Meeting of the Association for Computational Linguistics (ACL-COLING'98)*, Montréal, Canada, p. 768-774, 1998.

Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press., 2008.

McCarthy D., Navigli R., « The English lexical substitution task », *Language Resources and Evaluation*, vol. 43, n° 2, p. 139-159, 2009.

Metzler D., Croft W., « Combining the Language Model and Inference Network Approaches to Retrieval », *Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval*, vol. 40, n° 5, p. 735-750, 2004.

Mikolov T., Yih W.-t., Zweig G., « Linguistic Regularities in Continuous Space Word Representations », *2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL HLT 2013)*, Atlanta, Georgia, p. 746-751, 2013.

Miller G. A., « WordNet : An On-Line Lexical Database », *International Journal of Lexicography*, 1990.

Padó S., Lapata M., « Dependency-Based Construction of Semantic Space Models », *Computational Linguistics*, vol. 33, n° 2, p. 161-199, 2007.

Ponte J. M., Croft W. B., « A language modeling approach to information retrieval », *Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '98)*, p. 275-281, 1998.

Robertson S. E., Walker S., Hancock-Beaulieu M., « Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive », *Proc. of the 7th Text Retrieval Conference, TREC-7*, p. 199-210, 1998.

Ruiz-Casado M., Alfonseca E., Castells P., « Using context-window overlapping in Synonym Discovery and Ontology Extension », *Proceedings of RANLP-2005*, Borovets, Bulgarie, 2005.

Sahami M., Heilman T., « A web-based kernel function for measuring the similarity of short text snippets », *Proceedings of WWW'2006*, 2006.

Sahlgren M., « Vector-Based Semantic Analysis : Representing Word Meanings Based on Random Labels », *ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation*, Helsinki, Finland, 2001.

Strohman T., Metzler D., Turtle H., Croft W., Indri : A language-model based search engine for complex queries (extended version), Technical report, CIIR, 2005.

Turney P., « Mining the Web for Synonyms : PMIIR versus LSA on TOEFL », *Lecture Notes in Computer Science*, vol. 2167, p. 491–502, 2001.

Turney P., Pantel P., « From frequency to meaning : Vector space models of semantics », *Journal of Artificial Intelligence Research*, vol. 37, n° 1, p. 141-188, 2010.

Turtle H., Croft W., « Evaluation of an Inference Network-Based Retrieval Model », *ACM Transactions on Information System*, vol. 9, n° 3, p. 187-222, 1991.

Van de Cruys T., Mining for Meaning. The Extraction of Lexico-semantic Knowledge from Text, PhD thesis, University of Groningen, The Netherlands, 2010.

Van de Cruys T., Poibeau T., Korhonen A., « Latent vector weighting for word meaning in context », *in* A. for Computational Linguistics (ed.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 1012-1022, 2011.

Vechtomova O., Robertson S. E., « A Domain-Independent Approach to Finding Related Entities », *Information Processing and Management*, vol. 48, n° 4, p. 654-670, 2012.

Voorhees E. M., « Query Expansion Using Lexical-semantic Relations », *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, Springer-Verlag New York, Inc., New York, NY, USA, p. 61-69, 1994.

Ward G., « Moby Thesaurus », , Moby Project, 1996.

Yamamoto K., Asakura T., « Even Unassociated Features Can Improve Lexical Distributional Similarity », *Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010)*, Beijing, China, p. 32-39, 2010.

Zhitomirsky-Geffet M., Dagan I., « Bootstrapping Distributional Feature Vector Quality », *Computational Linguistics*, vol. 35, n° 3, p. 435-461, 2009.